ANNEX TO THE GLOSSARY

STATISTICAL CONCEPTS AND PROCEDURES

GA1. INTRODUCTION

The following brief presentation considers only those statistical terms important to an understanding of the material in this report. Similarly, the discussion of probability distributions focuses on those distributions that are most relevant to the discussions of uncertainty in the analyses presented in the main text. The goal is to present a discussion that is both (a) helpful to the reader who is unfamiliar with these topics and (b) sufficiently rigorous and consistent with standard technical definitions. Our choices of particular symbols to represent the quantities defined are not always consistent with those often found in statistical texts.

GA2. DEFINITIONS OF STATISTICAL TERMS

Definitions of some common statistical terms are presented in this section and some examples are given to illustrate the calculations. It is common to think of these definitions in the context of a set of measurements on a sample drawn from a population of objects whose true features are not known. The definitions given below are presented in this context. For example, we can measure the weights of a sample of 10-year-old children and use the information collected to estimate the mean and median weights for the population of children of that age. We can make such estimates, even though we will not ever know the true values for the group. In addition, we will be able to make some statements about the variability in our estimates.

Radiation doses and risks are not readily measurable quantities. We generally estimate the values of dose and risk using calculations and models that reflect the important processes and factors. In this report we use basic statistical terms to refer to sets of estimates of doses and risks that are obtained by calculation. We do not know the true value of a particular dose or risk but have central estimates of those quantities. We also have estimates of the range of likely values of dose and risk for the representative population groups considered.

Mean The mean value of a set of measurements of a quantity is the sum of the measured values divided by the number of measurements. The mean value is also often called the (arithmetic) average value. If we let *m* represent the mean of a number (n) of measurements x_1, x_2, x_3, . . . x_n, then

$$m = \frac{\sum_i x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

where the capital Greek letter sigma stands for the summation (addition) procedure and the letter i is an index that takes values 1, 2, 3, and so on. The true mean value is generally not known. We estimate it by making several *independent* measurements. An independent measurement is one whose outcome is not dependent upon or affected by any of the other measurements that are made. The mean may also refer to the average of independent estimates of a quantity, such as a dose or a risk, that can only be calculated.

Example The mean value of the set of eleven independent measurement values listed in Table GA-1 is:

$$m = \frac{12+9+5+7+13+6+2+3+8+12+5}{11} = \frac{82}{11} = 7.45$$

Table GA-1. Data Set for Examples, Eleven Independent Measurements

Measurement number	Value of measurement	Measurement number	Value of measurement	Measurement number	Value of measurement
1	12	5	13	9	8
2	9	6	6	10	12
3	5	7	2	11	5
4	7	8	3		

Median Of a set of n values, the median is the value that is as frequently exceeded (by other values in the set) as not. To determine the median value, arrange the set of values in order to generate the sequence: $x_1 \leq x_2 \leq x_3 \leq \ldots . \leq x_n$. When n is odd, the median is the value in the center of the sequence. The number of values that are smaller than the median is the same as the number of values that are larger than the median. If n is even, the median is not uniquely defined but is usually taken to be the average of the two central values. We also refer to the median of a set of estimates of a quantity, such as a dose or a risk, that can only be calculated.

Example The median of the set of eleven estimates in Table GA-1 is found by: (a) rearranging them in ascending order (2, 3, 5, 5, 6, 7, 8, 9, 12, 12, 13) and (b) choosing the central value (7) from the rearranged set. Five values are smaller than 7 and five exceed it. If the set contained the last ten values in Table GA-1, the rearranged sequence would be (2, 3, 5, 5, 6, 7, 8, 9, 12, 13). The median of this sequence is between 6 and 7 and, by convention, is taken to be 6.5. Again five values are smaller than the median and five values exceed it.

Mode The mode is the value that is measured (or estimated) most frequently. In the set of values in Table GA-1, there are two modes, 5 and 12. Distributions with two modes, called bimodal, are not uncommon in biological and environmental sciences.

Variance The variance of a set of measurements is the average value of the squares of the deviations of individual values from the mean value. The individual deviations from the mean are: $(x_1 - m)$, $(x_2 - m)$, $(x_3 - m)$, . . . $(x_n - m)$. When all the deviations are squared, added together, and divided by (n – 1), the result is the variance, usually denoted s^2. Mathematically, this is written as:

$$s^2 = \frac{\sum_i (x_i - m)^2}{n-1}$$

This particular equation defines the sample variance, an estimate of the true variance of the population from which the sample was drawn. The true variance of the population is generally not known.

Standard deviation The standard deviation (s) is the square root of the variance.

Coefficient of variation The coefficient of variation (cv) is the ratio of the standard deviation to the mean (s / m) and is often expressed as a percentage rather than a fraction.

Example The variance, standard deviation, and coefficient of variation of the example set of measurements in Table GA-1 are:

$$s^2 = \frac{(4.55)^2 + (1.55)^2 + (-2.45)^2 + (0.45)^2 + (5.55)^2 + (-1.45)^2 + (-5.45)^2 + (-4.45)^2 + (0.55)^2 + (4.55)^2 + (-2.45)^2}{10}$$

$$= \frac{20.7 + 2.4 + 6.0 + 0.2 + 30.8 + 2.1 + 29.7 + 19.8 + 0.3 + 20.7 + 6.0}{10} = \frac{137.8}{10} = 13.87$$

$$s = \sqrt{13.87} = 3.72$$

$$cv = \frac{s}{m} = \frac{3.72}{7.45} = 0.50 \text{ (a coefficient of variation of 50\%)}$$

Standard deviation of the mean The standard deviation of the mean (s_m) is the square root of the variance divided by the number of observations:

$$s_m = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

In our example, s_m = 3.72 / 3.32 = 1.12.

Geometric Mean The geometric mean (GM) of a set of n values is the nth root of the product of the values. To take a simple case, the geometric mean of (a, b) is the square root (second root) of a times b, which is written $(a \times b)^{1/2}$ or √(a x b). For the more general case, with values x_1, x_2, x_3, . . . x_n, the geometric mean is the nth root of the product of the n values.

$$GM = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}} = [(x_1)(x_2)(x_3)...(x_n)]^{\frac{1}{n}}$$

In the equation, the **Π** symbolizes the formation of the product. Alternative relationships that can be used to compute the geometric mean are:

$$\mu = \ln GM = \frac{\ln x_1 + \ln x_2 + \ln x_3 ... + \ln x_n}{n}$$

$$GM = e^{\mu}$$

That is, the natural logarithm (μ) of the geometric mean (GM) is the mean of the natural logarithms of the individual values and the GM is then e (the base of the natural logarithms) raised to the power μ.

Example The geometric mean of our previous set of eleven estimates in Table GA-1 is:

GM = [12 x 9 x 5 x 7 x 13 x 6 x 2 x 3 x 8 x 12 x 5]$^{1/11}$ = [849,139,200] $^{1/11}$ = 6.482

Alternatively,

$$\mu = \ln GM = \frac{\ln 12 + \ln 9 + \ln 5 + \ln 7 + \ln 13 + \ln 6 + \ln 2 + \ln 3 + \ln 8 + \ln 12 + \ln 5}{11}$$

$$\mu = \frac{2.485 + 2.197 + 1.609 + 1.946 + 2.565 + 1.792 + 0.693 + 1.099 + 2.079 + 2.485 + 1.609}{11}$$

$$\mu = \frac{20.560}{11} = 1.869$$

$$GM = e^{\mu} = e^{1.869} = 6.482$$

Geometric Standard Deviation The geometric standard deviation (*GSD*) reflects the deviations of the logarithms of the individual values (the x_i) from the logarithm of the geometric mean (μ). It is equal to e (the base of the natural logarithms) raised to the power σ, where σ is the standard deviation of the logarithms of the individual values.

$$\sigma = \sqrt{\frac{\sum_i (\ln x_i - \mu)^2}{n-1}}$$

$$GSD = e^{\sigma}$$

Thus, μ and σ, which are defined in terms of the logarithms of the x_i, are parallel quantities to the mean and standard deviation (m and s), which are defined in terms of the x_i themselves.

Example. The geometric standard deviation of our previous set of eleven estimates in Table GA-1 is found by first computing the differences between each of the ln x_i values and μ. All the ln x_i and μ have been computed above. The differences are then squared and used in the equation for σ:

$$\sigma = \sqrt{\frac{0.616^2 + 0.328^2 + (-0.260)^2 + (0.077)^2 + (0.696)^2 + (-0.077)^2 + (-1.176)^2 + (-0.770)^2 + 0.210^2 + 0.616^2 + (-0.260)^2}{10}}$$

$$\sigma = \sqrt{\frac{3.518}{10} = 0.593}$$

Then the GSD is computed: $GSD = e^{\sigma} = e^{0.593} = 1.8$.

Not all of the quantities defined above are used to describe all distributions. Some are more relevant than others for descriptions of particular distributions, as discussed below.

GA3. PROBABILITY DISTRIBUTIONS

There are many different probability distributions. Perhaps the most common mental model of a distribution of measured values in a sample is the normal distribution, the so-called “bell shaped curve” with which many people are familiar. In this section, we discuss the normal distribution and other distributions that are related to our risk estimation procedures and our calculations of the uncertainties associated with our risk estimates.

GA3.1 Normal Distribution

The normal distribution is an unbounded symmetric distribution, characterized by its mean (m) and standard deviation (s). Because the probability density function for the normal distribution is symmetric about the central value, the median is equal to the mean. The mode is indicated by the peak in the probability density function and is in this case also equal to the mean. Thus, the mean, median, and mode of a normal distribution are all the same. Values of the function f(x) outside the range of (m – 3s) to (m + 3s), 7 to 13 in this case, are quite small and are not shown clearly in Figure GA-1.

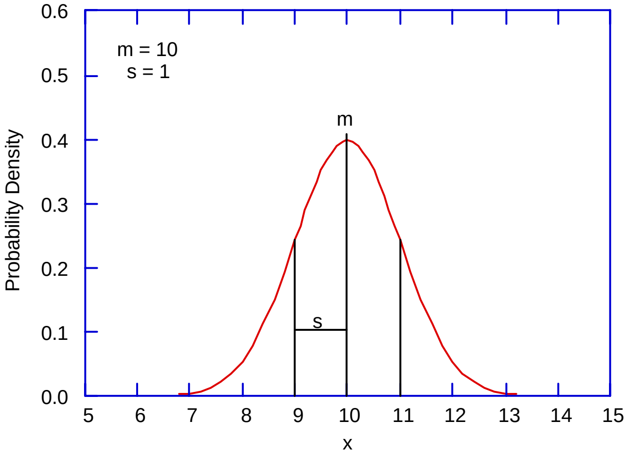

Figure GA-1. Probability density function for a normal distribution with a mean of 10 and a standard deviation of 1.

Figure GA-2 is a computer-generated histogram that approximates a normal distribution with a mean of 10 and a standard deviation of 2. The Monte Carlo procedure used to generate the figure involves independent random sampling of the idealized distribution to create a set of estimates that approximates the true distribution. It can be seen that the Monte Carlo procedure produced a good approximation but not a perfect normal distribution. The height of each bar is a measure of the probability that a particular value will fall within the range indicated by the width of the bar. The larger standard deviation implies a broader distribution and values between 4 and 16 are within three standard deviations of the mean.

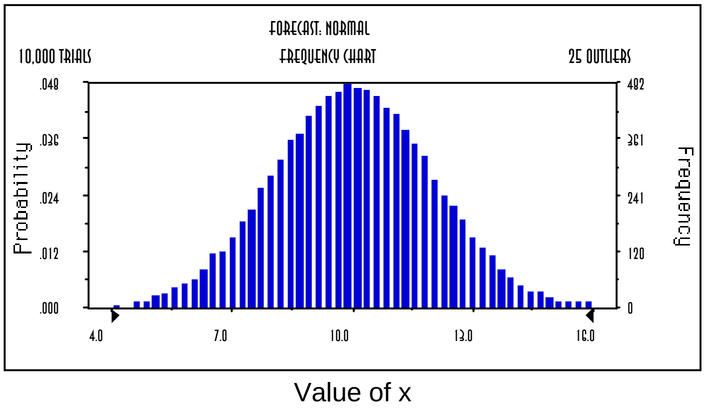

Figure GA-2. Histogram of distribution obtained by random sampling of a normal distribution with a mean of 10 and standard deviation of 2 using the Monte Carlo procedure used to evaluate uncertainties in the risk calculations.

A plot of the cumulative probability distribution provides another way to display the information shown above. Here the word cumulative implies a summation of the individual probabilities for the various values of x. Figure GA-3 shows such a plot for a simulation of a normal distribution with a mean of 10 and a standard deviation of 2. The cumulative probability increases as x increases because of the summation process. Although not readily seen from the plot, the minimum and maximum values were 2.0 and 16.8. The plot identifies three points on the distribution, the 25th percentile (cumulative probability = 0.25), the median, and the 75th percentile. One-fourth of the values are less than about 8.6 and one-fourth exceed 11.3; half must lie between those two points. The median is close to 10 as expected.

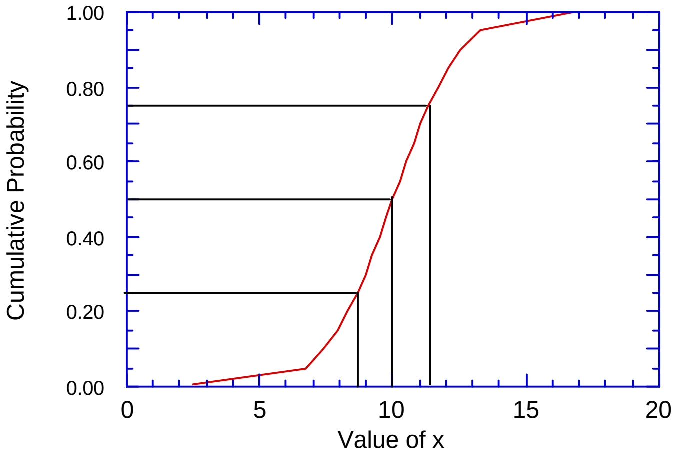

Figure GA-3. Cumulative probability plot for a normal distribution having a mean of 10 and a standard deviation of 2.

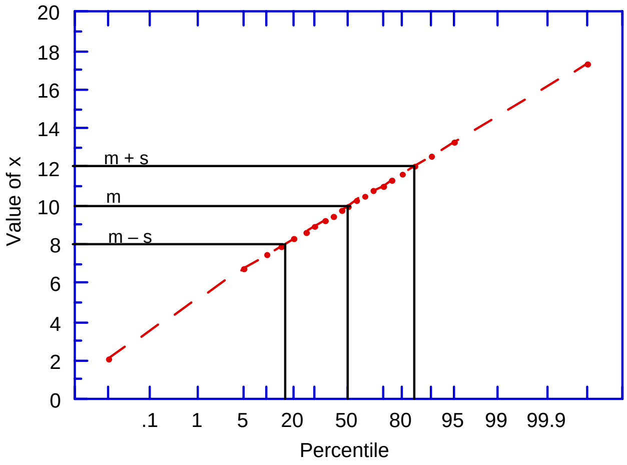

Figure GA-4. Cumulative probability plot for a normal distribution with a mean of 10 and a standard deviation of 2.

Figure GA-4 is another plot of the same information on a probability scale. In this figure, percentile markers for the distribution are plotted across the page. Values of x are plotted on the vertical axis and three of these are highlighted. The first is the mean and median of 10, which lies at the center or 50th percentile of the distribution. The diagram also illustrates that the value (m – s = 8) falls at the 16th percentile of the distribution, meaning that 16% of the values are less than 8. Similarly, the value of (m + s = 12) falls at the 84th percentile of the distribution. Sixteen percent of the values exceed (m + s). Thus, the zone between (m – s) and (m + s), between 8 and 12 in the example, contains 68% of all values.

GA3.2 Lognormal Distribution

A set of values whose logarithms are normally distributed is said to be lognormally distributed. The lognormal distribution is asymmetric and bounded at the lower end by zero; the upper end is unbounded. The distribution must be bounded below because the logarithm is defined only for positive numbers. Figure GA-5 shows a computer-generated approximation of a lognormal distribution that has a mean of 10 and a standard deviation of 4.

The fact that the lognormal probability distribution is asymmetric means that the mean and median will differ. The median of a lognormal distribution is always less than the mean of that distribution. In this case the median is 9.3 and the mode or most probable value is lower yet, 7.5. If plotted on log-probability paper, as in Figure GA-6, the cumulative probability distribution for a lognormal distribution consists of a straight line. Figure GA-6 is similar to Figure GA-4 above in that the percentiles are across the page and the values are on the vertical axis. The main difference between the plots is that the vertical scale in Figure GA-6 is logarithmic with major divisions that are factors of ten apart.

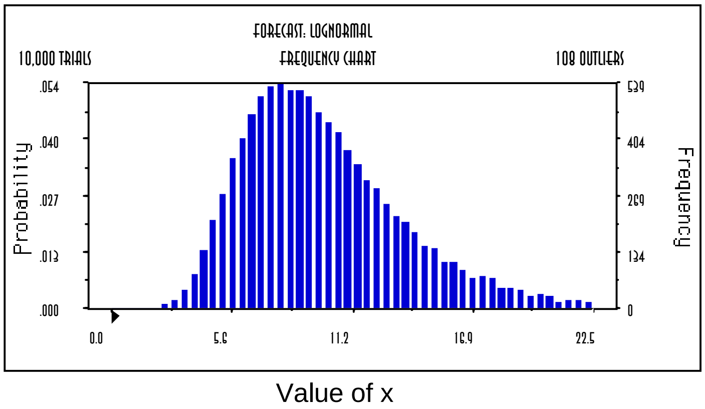

Figure GA-5. Histogram of distribution obtained by random sampling of a lognormal distribution with a mean of 10 and standard deviation of 4 using the Monte Carlo procedure used to evaluate uncertainties in the risk calculations.

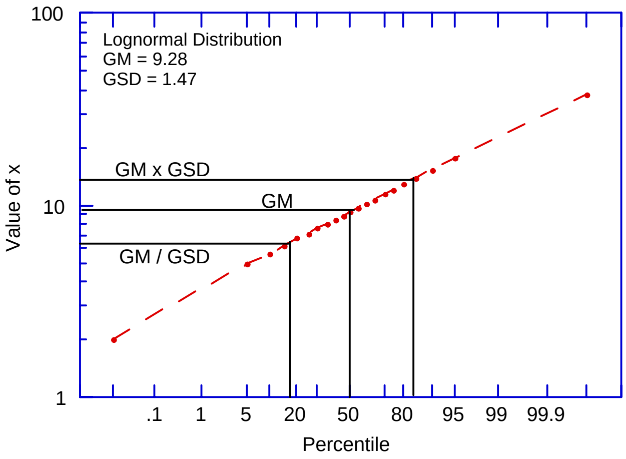

Figure GA-6. Cumulative probability plot for a lognormal distribution with a mean of 10 and a standard deviation of 4.

For this distribution, the central (50th percentile) value is the median or geometric mean and the spread is determined by the size of the geometric standard deviation. Definitions of these parameters were given above. The GSD of this example distribution is 1.5. The lower and upper bounds of the central 68% of the values are the quotient (GM ÷ GSD) and the product (GM x GSD). For this distribution those bounds are (9.3 ÷ 1.5 = 6.2) and (9.3 x 1.5 = 14), respectively.

There are direct relationships between the GM and GSD and the mean and standard deviation of a lognormal distribution. As noted above, μ = lnGM and σ = lnGSD. The mean (m) and variance (s^2) are computed using:

$$m = e^{\left(\mu + 0.5\sigma^2\right)}$$

$$s^2 = e^{\left(2\mu + \sigma^2\right)}\left(e^{\sigma^2} - 1\right)$$

and the standard deviation is the square root of the variance. The following relationships are used to compute the GM and GSD from the mean and standard deviation.

$$\mu = \ln\left[\frac{m}{\sqrt{1+\left(\frac{s}{m}\right)^2}}\right] \text{ and } GM = e^{\mu}$$

$$\sigma = \left[\sqrt{\ln\left[1+\left(\frac{s}{m}\right)^2\right]}\right] \text{ and } GSD = e^{\sigma}$$

GA3.3 Triangular Distribution

The triangular distribution is a bounded probability distribution that may be either asymmetric or symmetric. Examples of the probability density functions for two triangular distributions are shown in Figure GA-7. In both, the values of the distribution are shown along the x-axis while the probability density, f(x), is indicated by the height of the line above that value. The distribution on the left is asymmetric, with the mode (*b*) close to the lower bound (*a*) and far from the upper bound (*c*). The triangular distribution on the right is symmetric about the mode, with *a* and *c* equidistant from *b*.

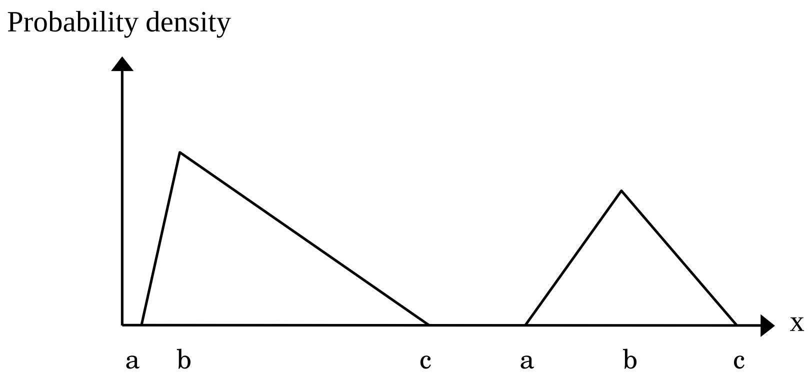

Figure GA-7. Probability density functions for two triangular distributions.

The histogram of a computer-generated triangular distribution is shown below in Figure GA-8. (It can be seen that the statistical procedure used to generate the distribution did not produce a perfect triangle.) This example is asymmetric, with a lower bound of a = 9, a mode of b = 10, and an upper bound of c =17.

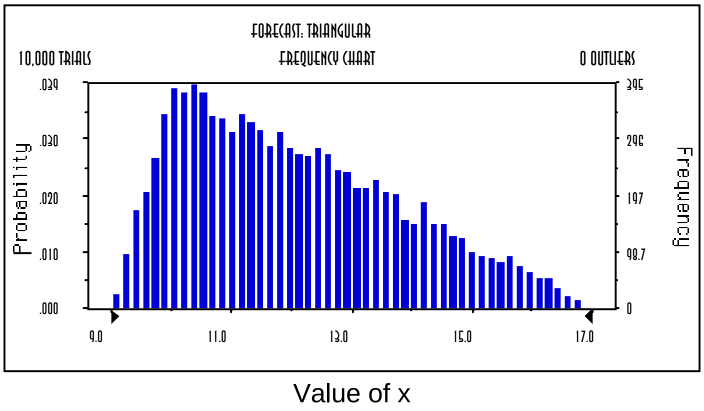

Figure GA-8. Histogram of distribution obtained by random sampling of a triangular distribution with a lower bound of 9, a mode of 10, and an upper bound of 17 using the Monte Carlo procedure used to evaluate uncertainties in the risk calculations.

The mean value of and variance for a triangular distribution are

$$m = \frac{a+b+c}{3}$$

$$s^2 = \frac{a(a-b)+b(b-c)+c(c-a)}{18}$$

The median of a triangular distribution is:

$$a + \sqrt{\frac{(c-a)(b-a)}{2}} \quad \text{if } b \geq \frac{a+c}{2}$$

$$c - \sqrt{\frac{(c-b)(c-a)}{2}} \quad \text{if } b \leq \frac{a+c}{2}$$

The mean, variance, standard deviation, and median of the distribution in Figure GA-8 are:

$$m = \frac{9+10+17}{3} = 12$$

$$s^2 = \frac{9(9-10)+10(10-17)+17(17-9)}{18} = \frac{-9-70+136}{18} = 3.17$$

$$s = \sqrt{3.17} = 1.78$$

$$\text{median} = 17 - \sqrt{\frac{(17-10)(17-9)}{2}} = 17 - 5.29 = 11.7$$

A cumulative probability plot for the example triangular distribution is shown in Figure GA-9. Like Figure GA-4, this plot utilizes a linear vertical axis and a horizontal probability scale. The median value of the distribution is highlighted in the figure. Values for other percentiles of interest can be read from the plot.

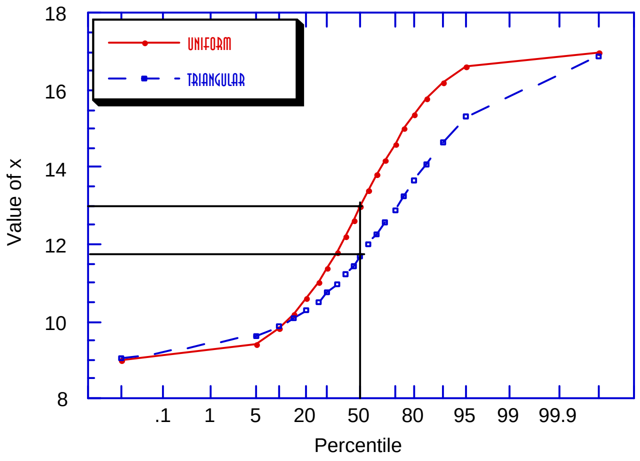

Figure GA-9. Cumulative probability plots for a triangular distribution with a mode of 10 and bounds of 9 and 17 and for a uniform distribution with the same bounds.

GA3.4 Uniform Distribution

The uniform probability distribution assigns equal likelihood to all values within the range of the distribution. An example of the probability density function for a uniform distribution, with lower bound a and upper bound b, is shown in Figure GA-10.

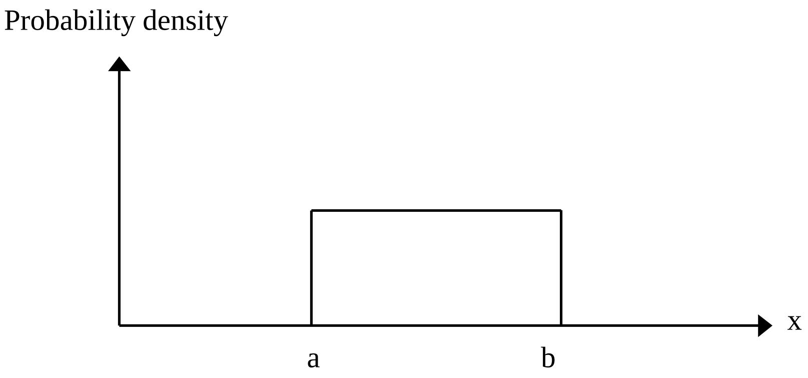

Figure GA-10. Probability density function for a uniform distribution.

For the uniform distribution, the mean and variance are:

$$m = \frac{a+b}{2}$$

$$s^2 = \frac{(b-a)^2}{12}$$

Because the distribution is symmetric, the median is the same as the mean. The mode of a uniform distribution is not uniquely defined.

A computer-generated uniform distribution with bounds of 9 and 17 is shown in Figure GA-11. A plot of the cumulative probability for that uniform distribution was shown in Figure GA-9 and the difference between the uniform and triangular distributions with the same bounds can be seen.

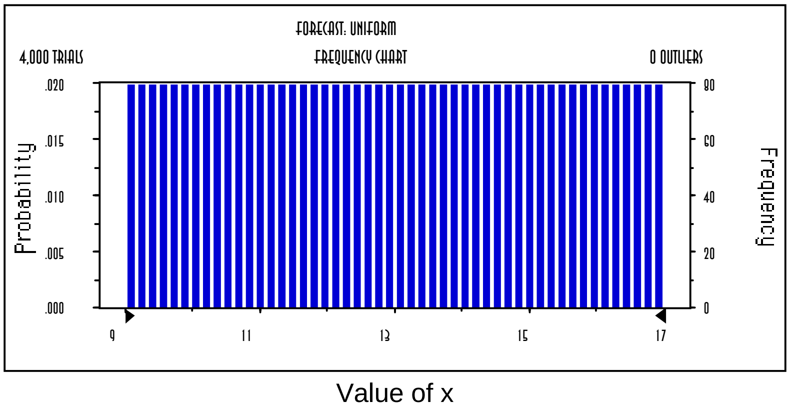

Figure GA-11. Histogram of distribution obtained by random sampling of a uniform distribution with a lower bound of 9 and an upper bound of 17 using the Monte Carlo procedure used to evaluate uncertainties in the risk calculations.

GA4. COMBINING SOURCES OF UNCERTAINTY

In some cases it is simple to determine the uncertainty in a result that involves multiple components each of which is itself uncertain. There are explicit formulas to perform the calculations for those cases. In other cases, the calculation of combined uncertainty must be performed using a Monte Carlo approach. In this section, formulas for the simple cases are given first. Then an example of a complex Monte Carlo calculation is presented.

GA4.1 Simple Combinations

We consider two quantities with mean values *Q* and *R* and, to avoid subscripts, we call their standard deviations *q* and *r*, respectively. The following expressions for the uncertainty in a third quantity of *P* that is derived from *Q* and *R* assume that *Q* and *R* are independently determined and that their values are not correlated with one another.

For the sum $P = Q + R$ or the difference $P = Q – R$, the variance of P is equal to the sum of the variances of Q and R. The standard deviation (p) is equal to the square root of the variance, hence:

$$p = \sqrt{p^2} = \sqrt{q^2 + r^2}$$

For a product or quotient, the square of the fractional uncertainty in P is equal to the sum of the squares of the fractional uncertainties of the components.

$$\frac{p}{P} = \sqrt{\left(\frac{p}{P}\right)^2} = \sqrt{\left(\frac{q}{Q}\right)^2 + \left(\frac{r}{R}\right)^2}$$

These relationships have great utility and there are more complex equations for situations when the variables Q and R are correlated. However, in environmental and biological sciences it is often found that uncertainty distributions are irregular or lognormal and those are not well characterized by the mean and standard deviation.

GA4.2 Uncertainties in Simple Multiplicative Models Whose Variables Have Lognormal Uncertainty Distributions

The generalized equations for dose or risk estimation are examples of calculations that might involve several lognormal distributions. We can write a simplified general equation for cancer risk (R) from a radionuclide release as:

$$R = Q \text{ x } D \text{ x } T \text{ x } A \text{ x } F$$

in which Q is the source term or amount of radionuclide release, D is a term that reflects dispersion and dilution of the radioactivity in air or water, the term T reflects all the transfer processes (such as deposition, animal metabolism, bioconcentration factors, and so on) that affect the human intake of radioactivity that results, the term A gives the absorbed dose to tissue of concern per unit intake, and F is a factor that converts tissue dose to cancer risk.

If the uncertainty distributions of Q, D, T, A, and F are all lognormal and independent, then the uncertainty distribution of the product R can be estimated using a simple arithmetic procedure. This pleasing simplification is due to the definition and properties of logarithms and the method for combining uncertainties of sums given above. We recall that μ_Q is the logarithm of the geometric mean (GM) or median of the distribution for Q, and σ_Q is the logarithm of the geometric standard deviation (GSD) of the distribution for Q, and use similar terminology for the other distributions. The central estimate of the distribution of risk is

$$GM_R = e^{\mu_R} \text{ where } \mu_R \text{ is given by: } \mu_R = \mu_Q + \mu_D + \mu_T + \mu_A + \mu_F$$

and the corresponding geometric standard deviation is

$GSD_R = e^{\sigma_R}$ where σ_R is given by: $\sigma_R = \sqrt{\left[\sigma_Q{}^2 + \sigma_D{}^2 + \sigma_T{}^2 + \sigma_A{}^2 + \sigma_F{}^2\right]}$

GA4.3 Use of Monte Carlo Calculations are Used to Propagate Uncertainties

Many calculations are not as straightforward as those discussed above. The uncertainty distributions may differ from lognormal and the calculational procedure may not be one that involves simple multiplication and division. It should be noted that when the distributions are lognormal, carrying uncertainties through additions and subtractions is in fact more complex than the process described above for multiplication.

It is nonetheless desirable to obtain not only a reliable central estimate of the result but to simultaneously define the uncertainty bounds associated with that estimate. Monte Carlo methods, which rely on random sampling of the uncertainty distributions and repeated calculations of model output values, can be used to achieve this goal.

Consider our previous model in which R is the product of the parameters *Q*, *D*, *T*, *A*, and *F*. The Monte Carlo procedure can be applied regardless of the types of uncertainty distributions for those parameters. Briefly, the procedure involves sampling independently from the five parameter distributions to obtain values Q_1, D_1, T_1, A_1, and F_1 and using these to compute an estimate of the risk, R_1. The procedure is repeated, using Q_2, D_2, T_2, A_2, and F_2, and to compute a second estimate, R_2, and so on until a large number of estimates of *R* have been made. We then examine the distribution of these estimates of *R* and find out what the central value is and what the appropriate uncertainty bounds for that estimate are. This procedure carries the various uncertainties in the input parameters forward and they are reflected in the resulting distribution of *R*.

As a simple example of the procedure, we consider a release of plutonium oxide to the atmosphere under steady meteorological conditions and compute the lung cancer risk for a representative individual who remains in the path of the plume throughout the several hour release period. In this example the term T is simple because it is only required to reflect the breathing rate. Table GA-2 summarizes the parameter distributions used for the calculation.

Table GA-2. Uncertainty Distributions for Input Variables for Example Monte Carlo Calculation

Input variable and units	Description	Type of distribution	Distribution parameters
Q (µCi ^{239}Pu)	Quantity released	Uniform	a = 50, b = 250
D (s m^{-3})	Atmospheric dispersion factor at point of exposure	Triangular	a = 1 x 10^{-7}, b = 5 x 10^{-7}, c = 3 x 10^{-6}
T (m^3 s^{-1})	Breathing rate while exposed	Lognormal	GM = 5.0 x 10^{-4}, GSD = 1.42
A (rad µCi^{-1})	Absorbed dose per unit intake	Lognormal	GM = 16, GSD = 1.72
F (risk rad^{-1})	Lung cancer risk coefficient	Triangular	a = 1 x 10^{-4}, b = 5 x 10^{-4} , c = 9 x 10^{-4}

The distributions described in Table GA-2 are illustrated in Figure GA-12. The plots shown there are the idealized probability density functions. As shown in previous examples, the Monte Carlo realizations are not perfect replicates of the underlying distributions.

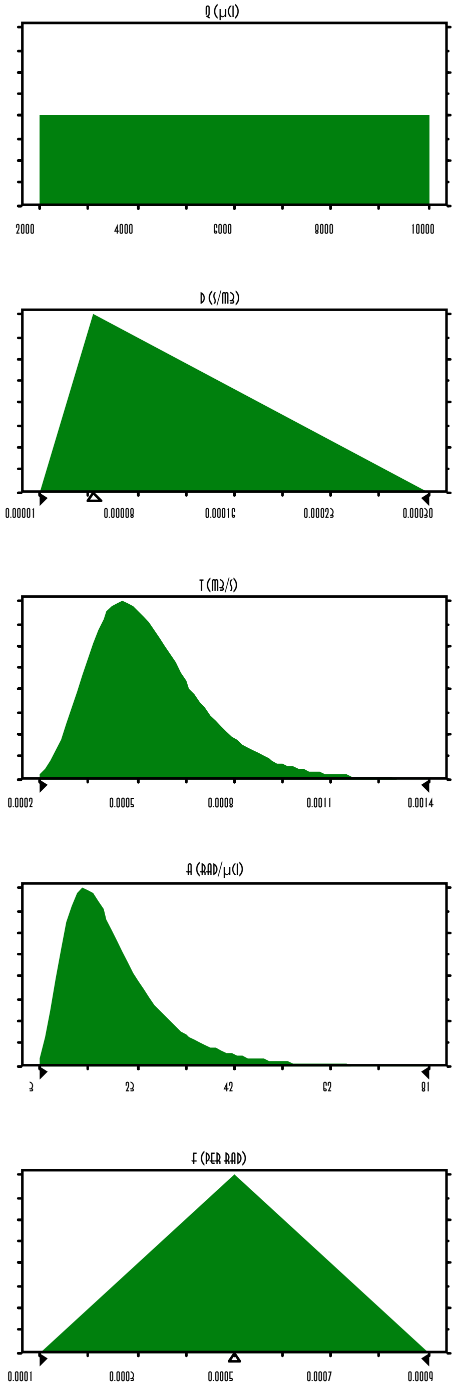

Figure GA-12. Distributions of the input variables used in the example Monte Carlo calculation of lung cancer risk. Distribution parameters are also given in the table.

Figure GA-13 shows the frequency distribution of the results of the example calculation of lung cancer risk. The shape of the distribution appears to be approximately lognormal with a relatively long tail. We see that roughly 0.8% of the values lie outside the plotted range. The overall range of risk estimates is from less than 4 x 10^{-8} (0.00000004) to nearly 6 x 10^{-5} (0.000006).

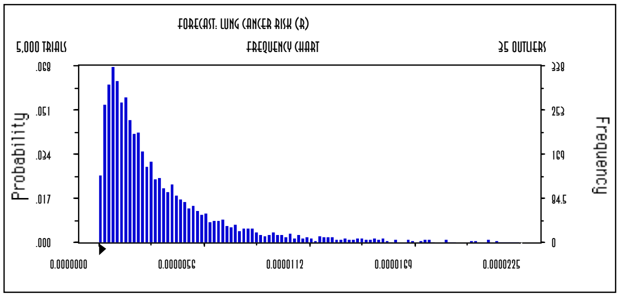

Figure GA-13. Frequency distribution of estimates of lung cancer risk for the example Monte Carlo calculation.

The range of estimates is shown more clearly in the plot of the cumulative probability distribution of the risk estimates (Figure GA-14). The fact that the line representing the cumulative distribution is nearly straight on this log-probability plot confirms that the risk estimates are approximately lognormally distributed. In the figure, the median value of 2.2 x 10^{-6} is highlighted as are the 5th and 95th percentile values. The geometric standard deviation of the distribution is about 3, indicating a broad range of estimates.

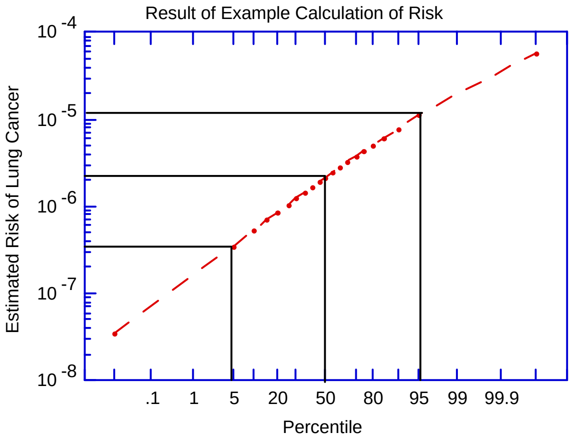

Figure GA-14. Cumulative probability distribution of estimates of lung cancer risk generated in the example Monte Carlo calculation.

The 5th to 95th percentile range is often used in expressions of the uncertainty bounds for the estimate. In this example, the 90 percent confidence interval for the lung cancer risk estimates is from 3.4 x 10^{-7} to 1.1 x 10^{-5}. Five percent of the estimates were less than 3.4 x 10^{-7} and five percent exceeded 1.1 x 10^{-5}. Other percentile values can be read from the curve; for example, one percent of the estimates exceeded 2 x 10^{-5}, the 99th percentile value.